# Determining factors behind the PageRank log-log plot

Yana Volkovich
University of Twente
Dept. of Applied Mathematics,
P.O. Box 217, 7500 AE
Enschede, The Netherlands
y.volkovich@ewi.utwente.nl

Nelly Litvak[*]
University of Twente
Dept. of Applied Mathematics,
P.O. Box 217, 7500 AE
Enschede, The Netherlands
n.litvak@ewi.utwente.nl

Debora Donato
Yahoo! Research
Barcelona Ocata 1, 1st floor
08003
Barcelona Catalunya, Spain
debora@yahoo-inc.com

## ABSTRACT

We study the relation between PageRank and other parameters of information networks such as in-degree, out-degree, and the fraction of dangling nodes. We model this relation through a stochastic equation inspired by the original definition of PageRank. Further, we use the theory of regular variation to prove that PageRank and in-degree follow power laws with the same exponent. The difference between these two power laws is in a multiple coefficient, which depends mainly on the fraction of dangling nodes, average in-degree, the power law exponent, and damping factor. The out-degree distribution has a minor effect, which we explicitly quantify. Our theoretical predictions show a good agreement with experimental data on three different samples of the Web.

## Keywords

PageRank, Power law, Recursive stochastic equations, Regular variation, Web graph

## MSC 2000

90B15, 68P10, 60J80

## 1. INTRODUCTION

Originally created for Web ranking, *PageRank* has become a major method for evaluating popularity of nodes in information networks. Besides its primary application in search engines, PageRank is successfully used for solving other important problems such as spam detection [20], graph partitioning [5], and finding gems in scientific citations [15], just to name a few. The PageRank [12] is defined as a stationary distribution of a random walk on a set of Web pages. At each step, with probability $c$, the random walk follows a randomly chosen outgoing link, and with probability $1-c$, the walk starts afresh from a page chosen at random according to some distribution $f$. Such random jump also occurs if a page is *dangling*, i.e. it does not have outgoing links. In the original definition, the teleportation distribution $f$ is uniform over all Web pages. Then the PageRank values satisfy the equation

$$PR(i) = c\sum_{j\to i}\frac{1}{d_j}PR(j) + \frac{c}{n}\sum_{j\in\mathcal{D}}PR(j) + \frac{1-c}{n},\ i=1,\ldots,n, \tag{1}$$

where $PR(i)$ is the PageRank of page $i$, $d_j$ is the number of outgoing links of page $j$, the sum is taken over all pages $j$ that link to page $i$, $\mathcal{D}$ is a set of dangling nodes, $n$ is the number of pages in the Web, and $c$ is the damping factor, which is a constant between 0 and 1.

From equation (1) it is clear that the PageRank of a page depends on popularity and the number of pages that link to it. Thus, it can be expected that the distribution of PageRank should be related to the distribution of *in-degree*, the number of incoming links. Most of experimental studies of the Web agree that in-degree follows a power law with exponent $\alpha = 1.1$ for cumulative plot, which corresponds to the famous value 2.1 for the density. Pandurangan et al. [27] discovered that PageRank also follows a power law with the same exponent. Further experiments [9, 16, 18] confirmed this phenomenon. Mathematical justifications have been proposed in [6, 19] for the preferential attachment models [3], and in [24], where the relation between PageRank and in-degree is modeled through a stochastic equation.

At this point, it is important to realize that PageRank is a *global* characteristic of the Web, which depends on in-degrees, *out-degrees*, correlations, and other characteristics of the underlying graph. In contrast to in-degrees, whose impact on the PageRank log-log plot is thoroughly explored and relatively well understood, the influence of out-degrees and dangling nodes has hardly received any attention in the literature. It is however a common belief that dangling nodes are important [17] whereas out-degrees (almost) do not affect the PageRank [18]. We also note that in the literature, there is no common agreement on the out-degree distribution. On the Web data, Broder et al. [13] report a power law with exponent about 2.6 for the density, whereas e.g. Donato et al. [16] obtain a distribution, which is clearly not a power law. On the other hand, for Wikipedia [14], out-degree seems to follow a power law with the same exponent as in-degree.

In the present paper we investigate the relations between PageRank and in/out-degrees, both analytically and exper-

[*]The work is supported by NWO Meervoud grant no. 632.002.401

imentally. Our analytical model is an extension of [24]. We view the PageRank of a random page as a random variable $R$ that depends on other factors through a stochastic equation resembling (1).

It is clear that the PageRank values in (1) scale as $1/n$ with the number of pages. In the analysis, it is more convenient to deal with corresponding *scale-free* PageRank scores

$$R(i) = nPR(i), \quad i = 1, \ldots, n, \tag{2}$$

assuming that $n$ goes to infinity. In this setting, it is easier to compare the probabilistic properties of PageRank and in/out-degrees, which are also scale-free. In the remainder of the paper, by PageRank we mean the scale-free PageRank scores (2). Then the original definition (1) can be written as

$$R(i) = c\sum_{j\to i}\frac{1}{d_j}R(j) + \frac{c}{n}\sum_{j\in\mathcal{D}}R(j) + 1 - c, \; i = 1, \ldots, n. \tag{3}$$

We are concerned with the *tail* probability $\mathbb{P}(R > x)$, i.e. the fraction of pages with PageRank greater than $x$, when $x$ is large. Our goal is to determine the asymptotic behavior of $\mathbb{P}(R > x)$, that is, we want to find a known function $r(x)$ such that $\mathbb{P}(R > x)/r(x) \to 1$ as $x \to \infty$. In this case, we say that $\mathbb{P}(R > x)$ and $r(x)$ are asymptotically equivalent, which essentially means that for large enough $x$, $\mathbb{P}(R > x)$ and $r(x)$ are close, and their log-log plots look the same. We formally describe power laws in terms of regular varying random variables, and we use recent results on regular variation to obtain the PageRank asymptotics. To this end, we provide a recurrent stochastic model for the power iteration algorithm commonly used in PageRank computations [23], and we obtain the PageRank asymptotics after each iteration.

The analytical results suggest that the PageRank and in-degree follow power laws with the same exponent. The out-degrees and dangling nodes affect only a multiple factor, for which we find an exact expression. It follows that the out-degree sequence has a truly minor influence whereas the fraction of dangling nodes has a slightly greater impact on the multiple coefficient. The experiments on the Indochina-2004 Web sample [1], on the EU-2005 Web sample [1], and on the Stanford Web [2], show that our model correctly predicts the evolution of the PageRank distribution through the series of power iterations, and it adequately captures the influence of the network parameters.

## 2. PRELIMINARIES

We start with preliminaries on the theory of regular variation, which is a natural formalization of power laws. More comprehensive details could be found, for instance, in [11]. We also refer to Jessen and Mikosch [22] for an excellent recent review.

*Definition 1.* A function $L(x)$ is *slowly varying* if for every $t > 0$,

$$\frac{L(tx)}{L(x)} \to 1 \quad \text{as} \quad x \to \infty.$$

*Definition 2.* A non-negative random variable $X$ is said to be *regularly varying* with index $\alpha$ if

$$\mathbb{P}(X > x) \sim x^{-\alpha}L(x) \quad \text{as} \quad x \to \infty, \tag{4}$$

for some positive slowly varying function $L(x)$.

Here, as in the remainder of this paper, the notation $a(x) \sim b(x)$ means that $a(x)/b(x) \to 1$.

The *asymptotic* equivalence (4) is a formalization of a power law. In words, it means that for large enough $x$, the tail distribution $\mathbb{P}(X > x)$ can be approximated by the regularly varying function $x^{-\alpha}L(x)$, which is, in turn, approximately proportional to $x^{-\alpha}$ due to the definition of $L$.

Regularly varying random variables represent a subclass of a much broader class of long-tailed random variables.

*Definition 3.* A random variable $X$ is *long-tailed* if for any $y > 0$,

$$\mathbb{P}(X > x + y) \sim \mathbb{P}(X > x) \quad \text{as} \quad x \to \infty. \tag{5}$$

Next lemma describes the behavior of a product and random sums of regular varying random variables. The relation (i) is known as Breiman's theorem (see e.g. Lemma 4.2.(1) in [22]). Properties (ii) and (iii) are, respectively, statements (2) and (5) of Lemma 3.7 in [22].

LEMMA 1. **(i)** *Assume that $X_1$ and $X_2$ are two independent non-negative random variables such that $X_1$ is regularly varying with index $\alpha$ and that $\mathbb{E}(X_2^{\alpha+\epsilon}) < \infty$ for some $\epsilon > 0$. Then*

$$\mathbb{P}(X_1X_2 > x) \sim \mathbb{E}(X_2^{\alpha})\mathbb{P}(X_1 > x).$$

**(ii)** *Assume that $N$ is regularly varying with index $\alpha \geq 0$; if $\alpha = 1$, then assume that $\mathbb{E}(N) < \infty$. Moreover, let $(X_i)$ be i.i.d. sequence such that $\mathbb{E}(X_1) < \infty$ and $\mathbb{P}(X_1 > x) = o(\mathbb{P}(N > x))$. Then as $x \to \infty$,*

$$\mathbb{P}\left(\sum_{i=1}^{N}X_i > x\right) \sim (\mathbb{E}(X_1))^{\alpha}\mathbb{P}(N > x).$$

**(iii)** *Assume that $\mathbb{P}(N > x) \sim r\mathbb{P}(X_1 > x)$ for some $r > 0$, that $X_1$ is regularly varying with index $\alpha \geq 1$, and $\mathbb{E}(X_1) < \infty$. Then*

$$\mathbb{P}\left(\sum_{i=1}^{N}X_i > x\right) \sim (\mathbb{E}(N) + r(\mathbb{E}(X_1))^{\alpha})\mathbb{P}(X_1 > x).$$

## 3. THE MODEL

### 3.1 In-degree

It is a common knowledge that in-degrees in the Web graph obey a power law with exponent about 2.1 for the density, which corresponds to 1.1 for cumulative plot. The power law exponent may deviate somewhat depending on a data set [8] and an estimator [26]. As in our previous work [24], we model the in-degree as an integer regularly varying random variable. To this end, we assume that the in-degree of a random page is distributed as $N(T)$, where $T$ is regularly varying with index $\alpha$ and $N(t)$ is the number of Poisson arrivals on the time interval $[0,t]$, when arrival rate is 1. If $T$ is regularly varying then $N(T)$ is also regularly varying and asymptotically identical to $T$ (see e.g. [24]). Thus, $N(T)$ is indeed integer and obeys the power law. To simplify the notation, we will use $N$ instead of $N(T)$ throughout the paper. The proposed formalization for the in-degree distribution allows us to model the number of terms in the summation in (3).

## 3.2 Out-degree and inspection paradox

Now, we want to model the weights $1/d_j$ in (3). Recall that $d_j$ is the out-degree of page $j$ that has a link to page $i$. In [24] we studied the relation between in-degree and PageRank assuming that out-degrees of all pages are constant, equal to the expected in-degree $d$. In this work, we make a step further allowing for random out-degrees.

We model out-degrees of pages linking to a randomly chosen page as independent and identically distributed random variables with arbitrary distribution. Thus, consider a random variable $D$, which represents the out-degree of a page that links to a particular randomly chosen page $i$. Note that $D$ is *not* the same random variable as an out-degree of a random page since the additional information that a page has a link to $i$, alters the out-degree distribution. This famous phenomenon, called *inspection paradox*, finds its mathematical explanations in Renewal Theory. The inspection paradox roughly states that an interval containing a random point tends to be larger than a randomly chosen interval [28]. For instance, in [29], a number of children in a family, to which a randomly chosen child belongs, is stochastically larger than a number of children in a randomly chosen family. Likewise, a number of out-links $D$ from a page containing a random link, should be stochastically larger than an out-degree of a random page. We will refer to $D$ as *effective out-degree*. The term is motivated by the fact that the distribution of $D$ is the one that participates in the PageRank formula.

Now, let $p_j$ be a fraction of pages with out-degree $j \geq 0$. Then we have

$$\lim_{n\to\infty} \mathbb{P}(D=j) = \frac{jp_j}{d}, \quad j \geq 1. \tag{6}$$

where $d$ is the average in/out-degree, and $n$ is the number of pages in the Web. For sufficiently large networks, we may assume that the distribution of $D$ equals to its limiting distribution defined by (6). Note that, naturally, the probability that a random link comes from a page with out-degree $j$ is proportional to $j$. This was implicitly observed by Fortunato et al. in [18], who in fact used (6) in their computations for the mean-filed approximation of PageRank.

## 3.3 Stochastic equation

We view the scale-free PageRank of a random page as a random variable $R$ with $\mathbb{E}(R) = 1$. Further, we assume that the PageRank of a random page does not depend on the fact whether the page is dangling. Indeed, it can be shown that the PageRank of a page can not be altered significantly by modifying outgoing links [7]. Moreover, experiments e.g. in [17] show that dangling nodes are often just regular pages whose links have not been crawled, for instance, because it was not allowed by `robot.txt`. Besides, even authentically dangling pages such as `.pdf` or `.ps` files, often contain important information and gain a high ranking independently of the fact that they do not have outgoing links. We note that such independence implies that the average PageRank of dangling nodes is 1, and thus the fraction of the total PageRank mass concentrated in dangling nodes, equals to the fraction of dangling nodes $p_0$:

$$p_0 = \frac{1}{n}\sum_{j\in\mathcal{D}} R(j).$$

Our goal is to model and analyze to which extent the tail probability $\mathbb{P}(R > x)$ for large enough $x$ depends on the in-degree $N$, the effective out-degree $D$, and the fraction of dangling nodes $p_0$. To this end, we model PageRank $R$ as a solution of a stochastic equation involving $N$ and $D$. Inspired by the original formula (3), the stochastic equation for the scale-free PageRank is as follows:

$$R \overset{d}{=} c\sum_{j=1}^{N}\frac{1}{D_j}R_j + [1-c(1-p_0)]. \tag{7}$$

Here $N$, $R_j$'s and $D_j$'s are independent; $R_j$'s are distributed as $R$, $D_j$'s are distributed as $D$, and $a \overset{d}{=} b$ means that $a$ and $b$ have the same probability distribution. As before, $c \in (0,1)$ is a damping factor.

We note that the independence assumption for PageRanks and effective out-degrees of pages linking to the same page, is obviously not true in general. However, there is also no direct relation between these values as there is no experimental evidence that such dependencies would crucially influence the PageRank distribution. Thus, we assume independence in this study.

The stochastic equation (7) is a generalization of the equation analyzed in [24], where it was assumed that $D_j$'s are constant. In order to demonstrate applicability of our model, we will use (7) to derive a mean-field approximation for the PageRank of a page with given in-degree. It follows from (6) that

$$\mathbb{E}\left(\frac{1}{D}\right) = \sum_{j=1}^{\infty}\frac{1}{j}\mathbb{P}(D=j) = \sum_{j=1}^{\infty}\frac{1}{j}\,\frac{jp_j}{d} = \frac{1-p_0}{d}.$$

Then, assuming that $\mathbb{E}(R_j) = 1$, $j = 1,2,\ldots$, we obtain

$$\mathbb{E}(R|N) = \frac{c(1-p_0)}{d}N + [1-c(1-p_0)]. \tag{8}$$

If $p_0 = 0$ then this coincides with the mean-field approximation by Fortunato et al. in [18], obtained directly from the PageRank definition under minimal independence assumptions and without considering dangling nodes.

Equation (7) belongs to the class of stochastic recursive equations that were discussed in detail in the recent survey by Aldous and Bandyopadhyay [4]. In particular, (7) has an apparent similarity with distributional equations motivated by branching processes and branching random walks. Such equations were studied in detail by Liu in [25] and his other papers. Taking expectations in (8), we see that if $\mathbb{E}(R_j) = 1$, $j = 1,2,\ldots$, then $\mathbb{E}(R)$ also equals 1. In Section 5 we will show that (7) has a unique solution $R$ such that $\mathbb{E}(R) = 1$.

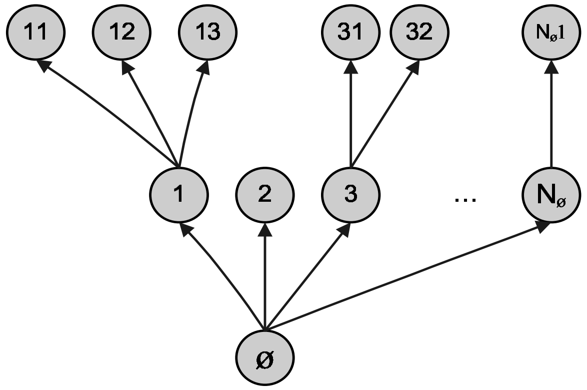


**Figure 1: An example of Galton-Watson tree**

## 4. MODEL FOR POWER ITERATIONS

In this section, we will introduce an iteration procedure for solving (7). This procedure can be seen as a stochastic model for the power iteration method commonly used in PageRank computations. We first present the notations, which are in lines with Liu [25].

Let $\left\{\left(N_u, \frac{1}{D_{u1}}, \frac{1}{D_{u2}}, \dots\right)\right\}_u$ be a family of independent copies of $\left(N, \frac{1}{D_1}, \frac{1}{D_2}, \dots\right)$ indexed by all finite sequences $u = u_1 \dots u_n$, $u_i \in \{1, 2, \dots\}$. And let $\mathbb{T}$ be the Galton-Watson tree with defining elements $\{N_u\}$ : we have $\emptyset \in \mathbb{T}$ and, if $u \in \mathbb{T}$ and $i \in \{1, 2, \dots\}$, then concatenation $ui \in \mathbb{T}$ if and only if $1 \le i \le N_u$. In other words, we indexed the nodes of the tree with root $\emptyset$ and the first level nodes $1, 2, ..N_\emptyset$, and at every subsequent level, the $i$th offspring of $u$ is named $ui$ (see Figure 1).

Now, we will iterate the equation (7). We start with initial distribution $R^{(0)}$, $\mathbb{E}\left(R^{(0)}\right) = 1$, and for every $k \ge 1$, we define the result of the $k$th iteration through a distributional identity

$$R^{(k)} \overset{d}{=} c \sum_{j=1}^{N} \frac{1}{D_j} R_j^{(k-1)} + [1 - c(1 - p_0)], \qquad (9)$$

where $N$, $R_j^{(k-1)}$ and $D_j$, $j \ge 1$, are independent. We argue that if $R^{(0)} \equiv 1$ then $R^{(k)}$ serves as a stochastic model for the result of the $k$th power iteration in standard PageRank computations. Indeed, according to (9) for $R^{(1)}$ we can obtain

$$R^{(1)} \overset{d}{=} c \sum_{j=1}^{N} \frac{1}{D_j} + [1 - c(1 - p_0)],$$

which clearly corresponds to the first power iteration with initial uniform vector:

$$PR^{(1)}(i) = c \sum_{j \to i} \frac{1}{d_j} + [1 - c(1 - p_0)],\ i = 1 \dots n.$$

This argument can be easily extended to further iterations.

Since PageRank vector is always a result of a finite number of iterations, it follows that $R^{(k)}$ describes the distribution of PageRank if the power iteration algorithm stops after $k$ steps. Assuming that in-degrees, effective out-degrees and $R_u^{(0)}$, $u \in \mathbb{T}$, are independent, and repeatedly applying (9), we derive the following representation for $R^{(k)}$:

$$R^{(k)} = c^k \sum_{u = u_1..u_k \in \mathbb{T}} \frac{1}{D_{u_1}} \cdots \frac{1}{D_{u_1..u_k}} R_{u_1..u_k}^{(0)} + [1 - c(1 - p_0)] \sum_{n=0}^{k-1} c^n Y^{(n)}, \quad k \ge 1, \qquad (10)$$

where

$$Y^{(n)} = \sum_{u = u_1 \dots u_n \in \mathbb{T}} \frac{1}{D_{u_1}} \cdots \frac{1}{D_{u_1 \dots u_n}},\ n \ge 1.$$

The random variable $Y^{(n)}$ represents the sum of the weights of the $n$th level of the Galton-Watson tree, where the root has weight 1, each edge has a random weight distributed as $1/D$, and the weight of a node is a product of weights of the edges, which are on the way from the root to this node.

In the subsequent analysis we will prove that iterations $R^{(k)}$, $k \ge 1$, converge to a unique solution of (7), and we will obtain the tail behavior of $R^{(k)}$ for each $k \ge 1$. This will give us the asymptotic behavior of the PageRank vector after an arbitrary number of power iterations.

## 5. ANALYTICAL RESULTS

First, we establish that our main stochastic equation (7) indeed defines a unique distribution $R$, that can serve as a model for the PageRank of a random page. The result is formally stated in the next theorem (the proof is given in Section 8).

THEOREM 1. *Equation (7) has a unique non-trivial solution with mean* 1 *given by*

$$R^{(\infty)} = \lim_{k \to \infty} R^{(k)} = [1 - c(1 - p_0)] \sum_{n=0}^{\infty} c^n Y^{(n)}. \qquad (11)$$

Now we are ready to describe the tail behavior of $R^{(k)}$, $k \ge 1$, which models the PageRank after $k$ power iterations. The main result is presented in Theorem 2 below.

THEOREM 2. *If* $\mathbb{P}\left(R^{(0)} > x\right) = o(\mathbb{P}(N > x))$, *then for all* $k \ge 1$,

$$\mathbb{P}(R^{(k)} > x) \sim C_k \mathbb{P}(N > x) \text{ as } x \to \infty,$$

*where* $C_k = \left(\frac{c(1-p_0)}{d}\right)^{\alpha} \sum_{j=0}^{k-1} c^{j\alpha} b^j$, *and* $b = d\mathbb{E}\left(1/D^{\alpha}\right) = \sum_{j=1}^{\infty} \frac{p_j}{j^{\alpha-1}}$.

The form of the coefficient $C_k$ arises from the proof, which relies on the results from [22]. The proof is provided in Section 8. For large enough $k$, $C_k$ can be approximated by

$$C = \lim_{k \to \infty} C_k = \frac{c^{\alpha}(1 - p_0)^{\alpha}}{d^{\alpha}(1 - c^{\alpha} b)}.$$

From the Jensen's inequality $\mathbb{E}(1/D^\alpha) \geq (\mathbb{E}(1/D))^\alpha$ and (3.3), it follows that $b \geq (1-p_0)^\alpha d^{1-\alpha}$, and hence,

$$C \geq \frac{c^\alpha (1-p_0)^\alpha}{d^\alpha (1 - c^\alpha (1-p_0)^\alpha d^{1-\alpha})}. \quad (12)$$

The last expression is the value of $C$ if out-degree of all non-dangling nodes is a constant. Note that if $\alpha \approx 1.1$, then the difference between the left- and the right-hand sides of (12) is really small for any reasonable out-degree distribution.

From Theorem 2 we can make interesting conclusions about the relation between PageRank and in/out-degrees. As it is commonly known from experiments, the power law exponent of the PageRank is the same as the power law exponent of in-degree. Clearly, this exponent is not affected by out-degrees. Thus, in-degree remains a major factor shaping the PageRank distribution. The multiple factor $C_k$, $k \geq 1$, depends mainly on the mean in-degree $d$, damping factor $c$, and the fraction of non-dangling nodes $(1-p_0)$. The values $p_j$, $j \geq 1$, that specify the out-degree distribution, have some effect on the coefficient $b$ but this results in a truly minor impact on the PageRank asymptotics. Hence, our results confirm the common idea that the out-degree distribution has a very little influence on the PageRank, but here we could also explicitly quantify this minor effect. In the next section we will compare out analytical findings with experimental results.

# 6. EXPERIMENTS

We performed experiments on Indochina-2004 and EU-2005 Web samples collected by The Laboratory for Web Algorithmics (LAW), Dipartimento di Scienze dell'Informazione (DSI) of the Universit degli studi di Milano [1]. We also used a Stanford-2002 Web sample [2]. In Figures 2–4 below we present cumulative log-log plots for in-degree/PageRank. The $y$-axis corresponds to the fraction of pages with in-degree/PageRank greater than the value on the $x$-axis. For in-degree, the power law exponent in evaluated using the maximum likelihood estimator from [26], and the straight line is fitted accordingly. For the PageRank, we plot the *theoretically predicted* straight lines obtained from Theorem 2.

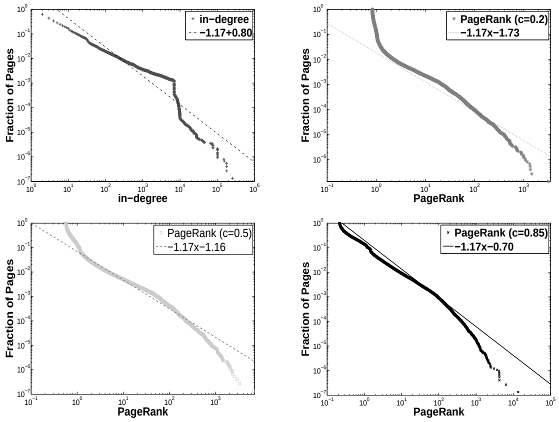
The Indochina set contains 7414866 nodes and 194109311 links. The results are presented in Figure 2 below. The in-degree plot resembles a power law except for the excessively large fraction of pages with in-degree about $10^4$. We suspect that this irregularity might be related to the specific crawling technique [10]. For more detail on this data set see [8]. For Indochina, we obtain a power law exponent 1.17 for cumulative plot, which is quite different from the result in [8]. This demonstrates the sensitivity of estimators for the power law exponent. Indeed, the exponent 0.6 in [8] reflects the behavior in the first part of the plot, whereas 1.17 gives more weight on the tail of the in-degree distribution.

We fit the straight line $y = -1.17x + 0.80$ into the in-degree plot and then compute the distance

$$\log_{10}(C) = \log_{10}\left(\frac{c^\alpha (1-p_0)^\alpha}{d^\alpha (1 - c^\alpha b)}\right)$$

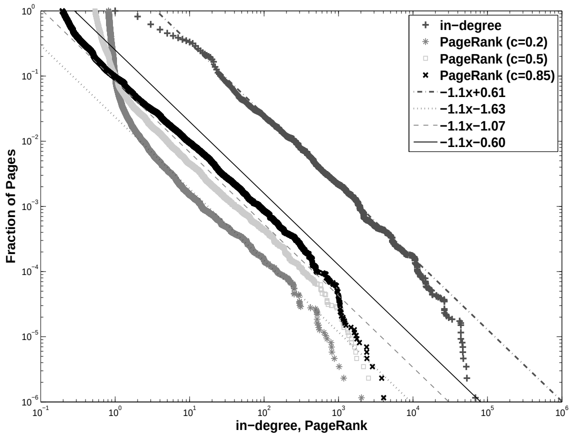
between the in-degree and the PageRank log-log plots for $c = 0.2, 0.5$, and 0.85. With $d = 26.17$, $p_0 = 0.18$, and $b = 0.65$, we obtain the following prediction for the PageRank log-log plot: $y = -1.17x - 1.73$ for $c = 0.2$, $y = -1.17x - 1.16$ for $c = 0.5$, and $y = -1.17x - 0.70$ for $c = 0.85$. In Figure 6 we show these *theoretically predicted* lines and the experimental PageRank log-log plots. We see that for this data set, our model provides the linear fit with a striking accuracy.

**Figure 2: Indochina data set: cumulative log-log plots for in-degree/PageRank. The straight lines for the PageRank plots are predicted by the model.**



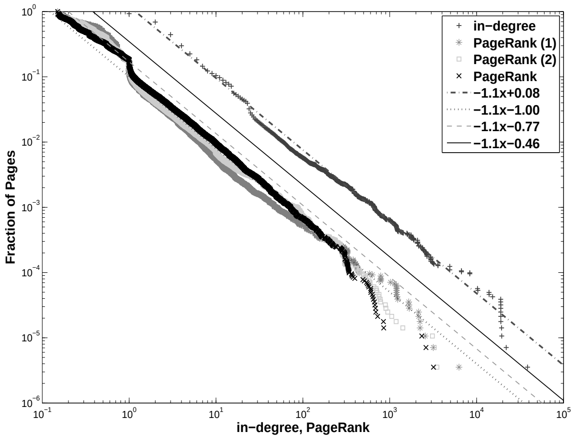
We performed the same experiment for EU-2005 of 862664 nodes and 19235140 links. In this data set in-degree shows a typical power law behavior, which is fitted perfectly by $y = -1.1x + 0.61$. We use the same approach to calculate the difference between the in-degree and PageRank plots for $d = 22.3$, $p_0 = 0.08$, $b = 0.70$. Thus, the theoretical prediction for the PageRank are $y = -1.1x - 1.63$, $y = -1.1x - 1.07$, and $y = -1.1x - 0.60$ for $c = 0.2, 0.5$, and 0.85, respectively. The log-log plots for experimental data, the fitted straight line for in-degree, and corresponding theoretical straight lines for PageRank, are presented in Figure 3.

**Figure 3: EU-2005 data set: cumulative log-log plots for in-degree/PageRank. The straight lines for the PageRank plots are predicted by the model.**



Finally, we verify out model for power iterations. For that, we use a smaller Web sample from [2] that contains 281903 pages and above 2.3 million links. In Figure 4 we show the cumulative log-log plot of in-degree, and the log-log plots of the PageRank after the 1st, the 2nd, and the last power iterations for the damping factor 0.85. To predict the difference between in-degree and PageRank's iterations we use the result of Theorem 2 for $d = 8.2032$, $p_0 = 0.006$, and $b = 0.8558$. Thus, if in-degree distribution could be fitted by $y = -1.1x + 0.08$, then $y = -1.1x - 1.00$, $y = -1.1x - 0.77$, and $y = -1.1x - 0.46$. are the predicted PageRank after the 1st, the 2nd, and the last power iterations, respectively. Although the obtained lines do not match perfectly the PageRank distribution, we see that our model correctly captures the dynamics of the PageRank distribution in successive power iterations. The difference between the theoretical prediction and the real data might occur because of the specific structure of this data set. For instance, the number of dangling nodes in this Web sample is negligibly small, which is not true for the real Web.

**Figure 4: Stanford data set: cumulative log-log plots for in-degree/PageRank. The straight lines for the PageRank plots are predicted by the model for the 1st, the 2nd, and the last power iterations.**



## 7. DISCUSSION

In this paper, we proposed an analytical stochastic model that helps to predict the shape of the PageRank log-log plot on basis of in-degree distribution, the damping factor, and the fraction of dangling nodes. It also follows form the model that the out-degree distribution has a truly minor impact on the PageRank. To make our mathematical model analytically tractable, we had to allow for several simplifying assumptions, such as independence of certain parameters and uniform teleportation. Experiments show that our theoretical model matches the Web data with a good accuracy.

One can argue that a uniform teleportation vector $f$ does not suit anymore for Web ranking [17]. Indeed, there are smarter choices of $f$ that take into account user's preferences, favor certain topics related to a query [21], or give higher weights to trusted pages for eliminating the spam [17]. The goal of this paper however was not improving the Web ranking but rather analyzing why the PageRank vector has certain properties reflected in its log-log plot. In order to capture the influence of in- and out-degrees, we had to make simplifying assumptions on other factors. However, we believe that our approach is promising in modeling relations between different parameters in the Web. In further research, we plan to gradually improve our model including dependencies, personalization, and other important factors relevant for the contemporary Web search.

## 8. PROOFS

Proof of Theorem 1. First, we establish that $R^{(\infty)}$ is well-defined random variable. We consider some initial distribution $R^{(0)}$ with $\mathbb{E}(R^{(0)}) = 1$. Then the first part of (10) has a mean $c^k(1-p_0)^k$, and hence it converges in probability to 0 because, by the Markov inequality, the probability that this term is greater than some $\epsilon > 0$ is at most $c^k(1-p_0)^k/\epsilon \to 0$ as $k \to \infty$. Further, since $(1-p_0)^{-n}Y^{(n)}$ is a martingale with mean 1, and $\lim_{n\to\infty}(1-p_0)^{-n}Y^{(n)}$ exists and it is finite (see [25]), the second part of (10) converges a.s. to $R^{(\infty)}$ as $k \to \infty$. It follows that (10) converges to $R^{(\infty)}$ in probability and according to the monotone convergence theorem

$$\mathbb{E}\left(R^{(\infty)}\right) = [1 - c(1-p_0)] \lim_{k\to\infty} \sum_{n=1}^{k} c^n \mathbb{E}\left(Y^{(n)}\right) = 1.$$

It is easy to verify that $R^{(\infty)}$ in (11) is a solution of (7). To prove the uniqueness, we assume that there is another solution with mean 1, then we take this solution as an initial distribution $R^{(0)}$ and repeat the argumentation above. Thus, we can conclude that there is no other fixed point of (7) with mean 1 except $R^{(\infty)}$. □

Proof of Theorem 2. We will use the induction. For $k = 1$, we derive

$$\begin{aligned}\mathbb{P}\left(R^{(1)} > x\right) &\sim \mathbb{P}\left(\sum_{j=1}^{N} \frac{c}{D_j} R_j^{(0)} + [1 - c(1-p_0)] > x\right)\\ &\sim \left(\frac{c(1-p_0)}{d}\right)^{\alpha} \mathbb{P}(N > x - [1 - c(1-p_0)])\\ &\sim C_1 \mathbb{P}(N > x) \text{ as } x \to \infty,\end{aligned}$$

where the second relation follows from Lemma 1$(ii)$ because $\mathbb{E}(N) = d < \infty$, $\mathbb{E}\left(R_1^{(0)}\right) = 1$, $\mathbb{E}\left(cD_1^{-1}R_1^{(0)}\right) = c(1-p_0)d^{-1} < \infty$, and $\mathbb{P}\left(cD_1^{-1}R_1^{(0)} > x\right) = o(\mathbb{P}(N > x))$, and the last relation follows from (5).

Now, assume that the result has been shown for $(k-1)$th iteration, $k \geq 2$. Then Lemma 1$(i)$ yields

$$\begin{aligned}\mathbb{P}\left(\frac{c}{D}R^{(k-1)} > x\right) &\sim c^{\alpha}\mathbb{E}\left(\frac{1}{D^{\alpha}}\right) C_{k-1}\mathbb{P}(N > x)\\ &= \frac{c^{\alpha}}{d}\, b\, C_{k-1}\mathbb{P}(N > x),\end{aligned}$$

where

$$\mathbb{E}\left(\frac{1}{D^{\alpha}}\right) = \sum_{j=1}^{\infty} \frac{p_j}{j^{\alpha}} = \frac{1}{d}\sum_{j=1}^{\infty} \frac{p_j}{j^{\alpha-1}} = \frac{1}{d}b.$$

Then, since $\mathbb{E}\left(cD^{-1}R^{(k-1)}\right) = c(1-p_0)d^{-1} < \infty$ and $\mathbb{E}(N) = d$, we apply Lemma 1(iii) to obtain

$$\begin{aligned}
\mathbb{P}(R^{(k)} > x) &\sim \mathbb{P}\left(\sum_{j=1}^{N} \frac{c}{D_j} R^{(k-1)} + [1 - c(1-p_0)] > x\right) \\
&\sim \left(c^\alpha b C_{k-1} + \left(\frac{c(1-p_0)}{d}\right)^\alpha\right) \mathbb{P}(N > x - [1 - c(1-p_0)]) \\
&\sim \left(c^\alpha b C_{k-1} + \left(\frac{c(1-p_0)}{d}\right)^\alpha\right) \mathbb{P}(N > x) \text{ as } x \to \infty,
\end{aligned}$$

for any $k \geq 2$. Here the last relation again follows from the property of long-tailed random variables (5).

Then for the constant $C_k$ we have

$$\begin{aligned}
C_k &= c^\alpha \; b \; C_{k-1} + \left(\frac{c(1-p_0)}{d}\right)^\alpha \\
&= \left(c^\alpha b \left(\frac{c(1-p_0)}{d}\right)^\alpha \sum_{j=0}^{k-2} c^{j\alpha} b^j + \left(\frac{c(1-p_0)}{d}\right)^\alpha\right) \\
&= \left(\frac{c(1-p_0)}{d}\right)^\alpha \sum_{j=0}^{k-1} c^{j\alpha} b^j .
\end{aligned}$$

□

## 9. REFERENCES

[1] http://law.dsi.unimi.it/. Accessed in January 2007.

[2] http://www.stanford.edu/∼sdkamvar/research.html. Accessed in March 2006.

[3] R. Albert and A. L. Barabàsi. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[4] D. J. Aldous and A. Bandyopadhyay. A survey of max-type recursive distributional equations. *Ann. Appl. Probab.*, 15:1047–1110, 2005.

[5] R. Andersen, F. Chung, and K. Lang. Local graph partitioning using pagerank vectors. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 475–486, Washington, DC, USA, 2006. IEEE Computer Society.

[6] K. Avrachenkov and D. Lebedev. PageRank of scale free growing networks. Technical Report 5858, INRIA, 2006.

[7] K. Avrachenkov and N. Litvak. The effect of new links on Google PageRank. *Stoch. Models*, 22(2):319–331, 2006.

[8] R. Baeza-Yates, C. Castillo, and E. Efthimiadis. Characterization of national Web domains. *To appear in ACM TOIT*, 2006.

[9] L. Becchetti and C. Castillo. The distribution of PageRank follows a power-law only for particular values of the damping factor. In *Proceedings of the 15th international conference on World Wide Web*, pages 941–942. ACM Press, New York, 2006.

[10] L. Becchetti, C. Castillo, D. Donato, and A. Fazzone. A comparison of sampling techniques for Web characterization. In *Workshop on Link Analysis (LinkKDD)*, 2006.

[11] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation.* Cambridge University Press, 1989.

[12] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Networks ISDN Systems*, 33:107–117, 1998.

[13] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Statac, A. Tomkins, and J. Wiener. Graph structure in the Web. *Comput. Networks*, 33:309–320, 2000.

[14] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardiand, and G. Caldarelli. Preferential attachment in the growth of social networks: the case of Wikipedia. Technical Report 0602026, arXiv/physics, 2006.

[15] P. Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with Google. Technical Report 0604130, arxiv/physics/, 2006.

[16] D. Donato, L. Laura, S. Leonardi, and S. Millozi. Large scale properties of the Webgraph. *Eur. Phys. J.*, 38:239–243, 2004.

[17] N. Eiron, K. S. McCurley, and J. A. Tomlin. Ranking the Web frontier. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 309–318, New York, NY, USA, 2004. ACM Press.

[18] S. Fortunato, M. Boguna, A. Flammini, and F.Menczer. How to make the top ten: Approximating PageRank from in-degree, 2005. arXiv.org/cs/cs.IR/0511016.

[19] S. Fortunato and A. Flammini. Random walks on directed networks: the case of PageRank, 2006. arxiv.org/physics/0604203.

[20] Z. Gyongyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *30th International Conference on Very Large Data Bases*, page 576587, 2004.

[21] T.H. Haveliwala. Topic-sensitive PageRank: A context-sensitive ranking algorithm for Web search. *IEEE TKDE*, 15(4):784–796, 2003.

[22] A. H. Jessen and T. Mikosch. Regularly varying functions. *Publications de L'Institut Mathematique, Nouvelle Série*, 79(93), 2006.

[23] A. N. Langville and C. D. Meyer. Deeper inside PageRank. *Internet Math.*, 1:335–380, 2003.

[24] N. Litvak, W. R. W. Scheinhardt, and Y. Volkovich. In-degree and PageRank: Why do they follow similar power laws? To appear in *Internet Math.*

[25] Q. Liu. Asymptotic properties and absolute continuity of laws stable by random weighted mean. *Stochastic Process. Appl.*, 95(1):83–107, September 2001.

[26] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46:323–351, 2005.

[27] G. Pandurangan, P. Raghavan, and E. Upfal. Using PageRank to characterize web structure. In *8th Annual International Computing and Combinatorics Conference (COCOON)*, Singapore, 2002.

[28] S. M. Ross. *Stochastic processes.* Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1996.

[29] S. M. Ross. The inspection paradox. *Probab. Engrg. Inform. Sci.*, 17:47–51, 2003.